The SNOMED Clinical Terms Development Process: Refinement and Analysis of Content

Amy Y. Wang, MD[1,2], Jeremiah H. Sable, MD, MHA[1], Kent A. Spackman, MD, PhD[1,2]

[1]College of American Pathologists, Northfield, IL
[2]Oregon Health & Science University, Portland, OR

SNOMED Clinical Terms® is a comprehensive concept-based health care terminology that was created by merging SNOMED RT® and Clinical Terms Version 3. Following the mapping of concepts and descriptions into a merged database, the terminology was further refined by adding new content, modeling the relationships of individual concepts, and reviewing the hierarchical structure. A quality control process was performed to ensure integrity of the data. Additional features such as subsets, qualifiers, and mappings to other coding systems were added or updated to facilitate usability. We then analyzed the content of the completed work. This paper describes the refinement processes and compares the actual content of SNOMED CT® with the early data obtained from analysis of the description mapping process. As predicted, the majority of concepts in SNOMED CT originated from SNOMED RT or CTV3, but not both.

INTRODUCTION AND BACKGROUND

The College of American Pathologists (CAP) and the National Health Service (NHS) in the United Kingdom formed a partnership in 1999 to merge their respective terminologies, SNOMED RT and Clinical Terms Version 3 (CTV3), into a comprehensive new work, titled SNOMED Clinical Terms or SNOMED CT.[†] The development of SNOMED CT was a collaborative effort that involved contributions from numerous individuals representing diverse scientific disciplines and healthcare organizations. The collaboration, planning, design, and early production have been described previously.[1]

An important phase of the SNOMED CT development process was the formation of description-to-concept maps between SNOMED RT and CTV3. A concept is a unit of thought, which is represented using one or more concept descriptions, or terms. A description-to-concept mapping process was used to merge the content of the two terminologies. SNOMED® clinical editors manually reviewed concepts for synonymy and assigned semantic equivalence, or "same-as" or supertype-subtype, or "is-a" maps between descriptions in one terminology and concepts in the other terminology. The pre-existing hierarchical or is-a relationships and attribute-value relationships of concepts were preserved. Ambiguous concepts were separated into multiple concepts, with the original concept codes retired from active use. (While inactive concepts are no longer used for coding new data, they remain in SNOMED CT with links to currently active concepts to accommodate any legacy data encoded using these concepts.) The result was a rough, merged terminology containing concepts, descriptions, and hierarchical relationships from SNOMED RT and CTV3.[2]

An analysis of the data from the early mapping process found that 28% of the description-to-concept maps between SNOMED RT and CTV3 were same-as maps, and 72% were is-a maps. Based on these findings, it was predicted that the majority of concepts in SNOMED CT would originate from either SNOMED RT or CTV3 alone, with a smaller number present in both. A number of concepts were newly created for SNOMED CT and thus were not contained in the source terminologies. After completion SNOMED CT First Release, we analyzed the content in order to determine how the actual composition of SNOMED CT compares with previous predictions based on mapping data.

DEVELOPMENT PROCESS

Editorial Guidance

The SNOMED International® Editorial Board determined the final specification of the content and structure of the SNOMED CT First Release. The editorial board prioritized the attributes to be modeled and released. Highest priority were those attributes that: 1) are understandable, reproducible, and useful; and 2) originated from SNOMED RT or CTV3 and were previously released. Table 1 shows the approved, defining attributes for SNOMED CT

[†] SNOMED CT and SNOMED RT are copyrighted works of the College of American Pathologists. Clinical Terms Version 3 was developed by the NHS and is a Crown copyright.

Disorder/Finding attributes

- Finding site
- Causative agent
- Associated morphology
- Episodicity
- Severity
- Onset
- Course

Procedure attributes

- Procedure site
- Method
- Direct morphology
- Direct substance
- Direct device
- Access
- Using
- Approach

Body structure attributes

- Part of
- Laterality

Table 1. Approved, defining attributes in SNOMED CT First Release.

First Release. Additional attributes are currently in development.

Addition of New Content

After the mapping process concluded and a merged database was formed, new concepts were added to the database from several sources. For example, sets of concepts used in cancer protocol checklists and clinical imaging were added. Concepts were also added to SNOMED CT representing generic clinical drugs (e.g., "Acetaminophen 325 mg tablet"). SNOMED RT had previously contained drug concepts only for generic ingredients (e.g., "Acetaminophen") and proprietary products (e.g., "Tylenol Extra Strength"). Concepts representing proprietary pharmaceutical products were not added to the core content of SNOMED CT, but will be included in extensions, which are discussed below.

Terminology Modeling

SNOMED clinical editors reviewed and modeled the concepts in the merged database using several tools. A Microsoft Access-based tool was developed in-house specifically for reviewing is-a relationships. SNOMED convent developers used this tool to review every pre-existing is-a relationship in the merged database for accuracy. Using a template-editing tool developed by the NHS, editors reviewed every attribute-value relationship for all concepts in the Finding, Disorder, and Procedure hierarchies. Individual concepts were modeled using the Apelon TDE, and the hierarchical relationships were then developed further using a description logic engine.

Quality Control

A quality review process was performed on the data to assess data integrity. We verified that every SNOMED RT and CTV3 concept was represented in SNOMED CT. The data was checked to ensure that every required field was populated with an appropriate value. A search was done for duplicate concepts, which were then merged together. The hierarchical relationships were also checked for internal consistency. For example, concepts that were subtypes of concepts from more than one top-level hierarchy (e.g., a concept is a subtype of "Disorders" and "Procedures") were detected and corrected.

Addition of Qualifiers

SNOMED authors in the UK used a custom-built tool to specify the qualifiers of concepts to allow further specification. For example, a disorder can be given allowable qualifiers of "Mild," "Moderate," and "Severe." When used with the "Severity" attribute, these values can provide more detailed coding of a disorder. The UK editors also specified allowable qualifying relationships that can be used to create more granular attribute-value relationships than those in SNOMED CT. For example, a user can specify "Bacterial endocarditis," which has "Causative agent: Bacterium" by creating a new relationship, "Causative agent: Streptococcus agalactiae" to represent a disorder that is not in SNOMED CT.

Mapping to Other Coding Systems

A one-to-one relationship exists between the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) morphology concepts and SNOMED CT tumor morphology concepts, and the concept codes in both are identical. Cross maps were performed between SNOMED CT concepts and categories in ICD-9-CM, ICD-10, and Office of Population, Censuses and Surveys Classification of Surgical Operations and Procedures, 4th Revision (OPCS-4). Concepts from the laboratory portion of Logical Observation Identifier Names and Codes (LOINC) were linked to SNOMED CT through hierarchical relationships. In addition, defining elements of LOINC concepts (e.g., component, property, method, timing) are represented by concepts in SNOMED CT.

Development of Subsets

A subset refers to a group of concepts, descriptions, or relationships in SNOMED CT that are appropriate for use in a particular language, dialect, country, specialty, organization, user or context. Although a comprehensive terminology is valuable, limiting the size and scope of available content can improve usability. For example, we have created subsets containing descriptions with American and British lexical variants. We developed a list of English words with corresponding American and British spellings to assign descriptions containing these variants to the appropriate subset. For example, "Anemia" was assigned to the US subset, while "Anaemia" was placed into the UK subset. A user in the US can hide items from the UK subset from view, and vice versa.

A non-human subset was created containing concepts exclusive to veterinary medicine.[3] This subset is not intended to include all concepts needed for veterinary medicine, because many concepts apply to both human and veterinary medicine. The non-human subset can be used to remove exclusively veterinary concepts from a user interface where only concepts relating to human medicine are needed.

Creation of Extensions

Extensions are sets of concepts that are created in accordance with the structure and authoring guidelines of SNOMED but are not edited, maintained, or distributed under the guidance of the SNOMED International Editorial Board. Developers, implementers, and users have the capability to create extensions containing specific concepts needed for a local environment.

Proprietary drugs and other products specific to the US or UK were originally represented in SNOMED RT and CTV3, respectively. These are not included in the SNOMED CT core content, but will contained in extensions. US proprietary drug concepts are contained in the US proprietary drug extension, which is included with the SNOMED CT First Release. An extension containing UK proprietary pharmaceuticals, medical supplies, and other brand-name products is currently under development. A second extension with UK specific legal and administrative concepts is also being developed. These UK-specific extensions will be included in upcoming SNOMED CT releases.

METHODS

We examined the concepts in the SNOMED CT First Release in order to determine the numbers and proportions of concepts originating from SNOMED RT, CTV3, both, or neither. We also determined the sources of concepts in each of the eighteen SNOMED CT top-level hierarchies. We then compared these findings to the earlier data analysis from the description mapping process.

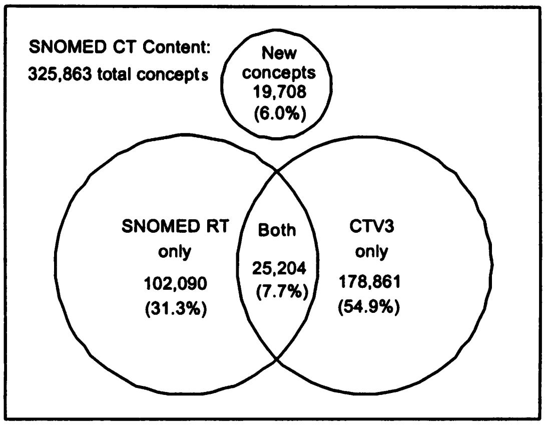

Figure 1. Sources of concepts in SNOMED CT.

RESULTS

SNOMED RT 1.1 and CTV3 September 2001, the source terminologies, contain 132,517 and 245,718 concepts respectively. SNOMED CT First Release has a total of 325,863 concepts. Of these, 102,090 (31.3%) originated in SNOMED RT only and 178,861 (54.9%) from CTV3 only. 25,204 (7.7%) concepts were from both terminologies and 19,708 (6.0%) concepts are new additions. These numbers and percentages are shown in Figure 1. Approximately 40,000 UK-specific proprietary and administrative concepts from CTV3 were not included in SNOMED CT and are being added to an extension. 5,781 proprietary drug concepts from SNOMED RT are now contained in the US propriety pharmaceutical extension and were not added to the SNOMED CT core content.

The numbers and percentages of concepts in each of the top-level hierarchies and the relative contribution from each source terminology are shown in Table 2. Figure 2 illustrates the relative contribution from the source terminologies to each area graphically.

DISCUSSION

The merging of SNOMED RT and CTV3 and subsequent refinement into SNOMED CT was a multidisciplinary, collaborative work. It required continual editorial guidance and the development of specialized tools. The majority of concepts were present in one or the other terminology, with a smaller number present in both. The approximately eight percent of concepts in SNOMED CT present in

Top-level Hierarchy	From SNOMED RT	From CTV3	From Both	New	Total
Disease	17,525 (24.6%)	42,687 (60.1%)	9,514 (13.4%)	1,282 (1.8%)	71,008
Procedure	17,334 (36.1%)	26,070 (54.2%)	3,319 (6.9%)	1,354 (2.8%)	48,077
Finding	5,076 (13.5%)	28,871 (76.8%)	1,529 (4.1%)	2,129 (5.7%)	37,605
Body structure	13,118 (42.3%)	9,634 (31.1%)	1,852 (6.0%)	6,389 (20.6%)	30,993
Substance	16,415 (50.9%)	8,002 (24.8%)	2,990 (9.3%)	4,854 (15.0%)	32,261
Organism	18,586 (76.9%)	964 (4.0%)	4,533 (18.7%)	95 (0.4%)	24,178
Qualifier value	2,454 (25.2%)	6,451 (66.3%)	449 (4.6%)	369 (3.8%)	9,723
Context category	3 (0.1%)	5,809 (99.9%)	2 (0.0%)	1 (0.0%)	5,815
Social context	1,983 (37.6%)	2,794 (53.0%)	476 (9.0%)	19 (0.4%)	5,272
Observable entity	100 (2.1%)	3,679 (76.2%)	127 (2.6%)	925 (19.1%)	4,831
Physical object	869 (28.3%)	1,767 (57.5%)	208 (6.8%)	229 (7.5%)	3,073
Staging and scales	129 (8.1%)	1,149 (72.2%)	39 (2.4%)	275 (17.3%)	1,592
Environments	8 (0.5%)	1,449 (99.1%)	5 (0.3%)	0 (0.0%)	1,462
Attribute	97 (10.1%)	825 (86.1%)	6 (0.7%)	30 (3.1%)	958
Specimen	362 (39.1%)	469 (50.7%)	82 (8.9%)	12 (1.3%)	925
Physical force	94 (46.5%)	80 (39.6%)	26 (12.9%)	2 (1.0%)	202
Events	30 (71.4%)	7 (16.7%)	5 (11.9%)	0 (0.0%)	42
Special concept	7,907 (16.5%)	38,154 (79.7%)	42 (0.1%)	1,743 (3.6%)	47,846
Total	102,090 (31.3%)	178,861 (54.9%)	25,204 (7.7%)	19,708 (6.0%)	325,863

Table 2. Numbers and percentages of concepts from SNOMED RT and CTV3 in the top-level hierarchies of SNOMED CT.

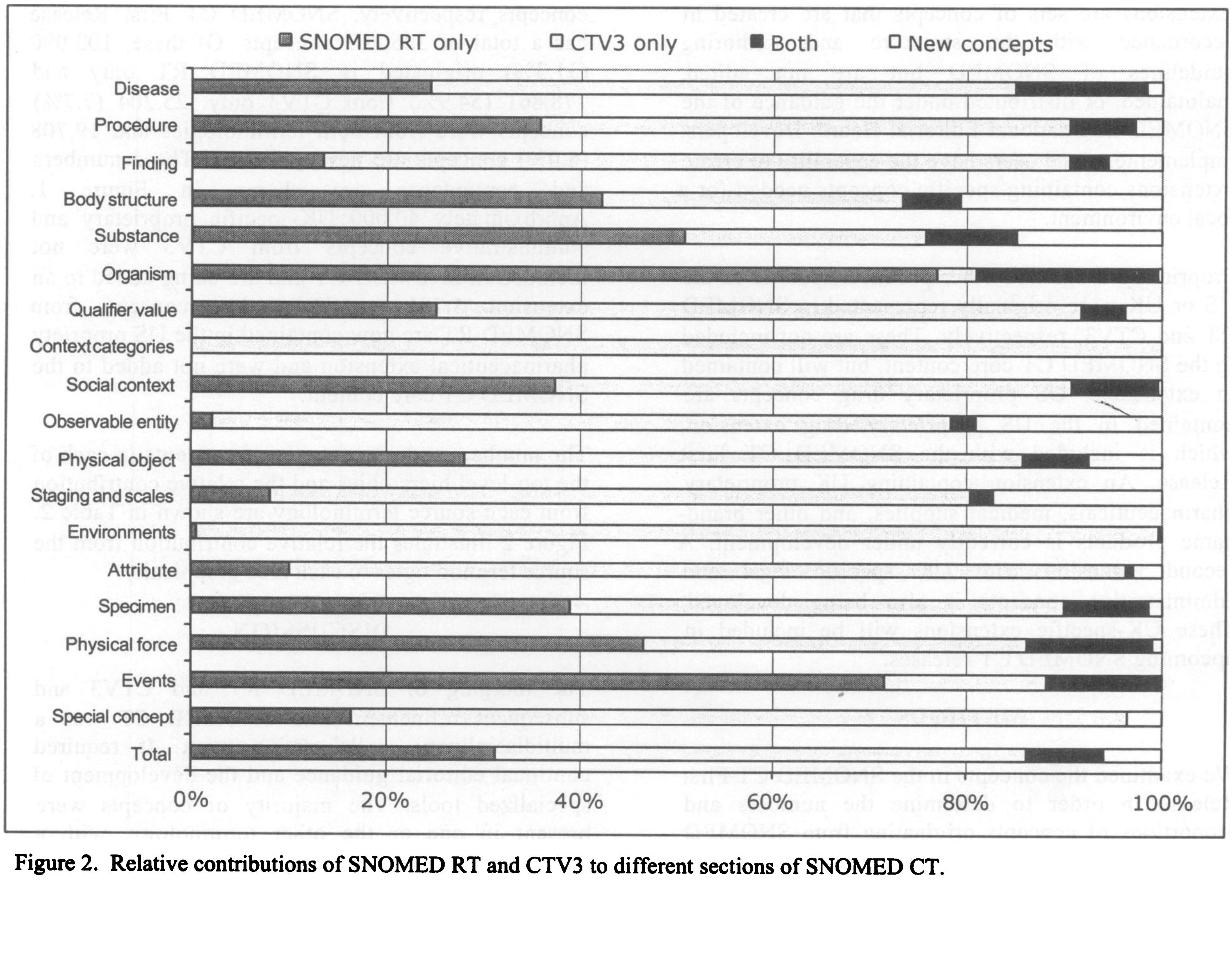

Figure 2. Relative contributions of SNOMED RT and CTV3 to different sections of SNOMED CT.

both SNOMED RT and CTV3 was somewhat lower than we expected, based on the description-to-concept same-as mapping rate of 28%.

It was challenging to predict the final composition of the content of SNOMED CT. One limitation of the previous estimate is that the analysis of the mapping data was based on description-to-concept maps and not concept-to-concept maps. Many duplicate concepts were merged and ambiguous concepts were separated, making it difficult to quantify the numbers of concepts from each source terminology. Another limitation of the mapping data is that authors focused their mapping efforts on the most clinically relevant portions of the source terminologies (i.e., findings, disorders, and procedures) and thus did not perform maps on all concepts and descriptions.

As predicted by the analysis of the mapping data, CTV3 contributed the majority of concepts in the Finding, Disorder, Procedure sections of SNOMED CT while SNOMED RT contained most of the concepts in the Body structure, Organism, Substance, and Morphology sections. These findings demonstrate that the relative contributions from SNOMED RT and CTV3 complement each other.

In addition to the core content, we have developed additional features to improve the functionality and usability of SNOMED CT. Qualifiers permit post-coordinated composition of concepts, making it possible to represent concepts not currently in SNOMED CT and preventing combinatorial explosion. Cross mapping facilitates the sharing of information that is coded using different systems. Subsets support the ability to limit the content visible in a user interface to those concepts most commonly used for a given situation. The ability to create extensions allows implementers to tailor SNOMED CT to their specific needs.

SNOMED CT development is an ongoing process that invites open discussion and feedback. Several enhancements being developed for upcoming releases include a UK proprietary product extension, UK administrative extension, primary care subset, and pathology subset. New concepts are being added in many areas, including nursing interventions, ophthalmology, and veterinary medicine.

CONCLUSION

The development of SNOMED CT was a complex project requiring the coordination of many individuals, organizations, resources, and processes. We have demonstrated that two comprehensive terminologies can be merged and developed into a unified work. Even more than originally predicted, the content of SNOMED CT is more comprehensive than the content of either SNOMED RT or CTV3 alone. In addition to the core content, SNOMED CT includes features that enhance usability and usefulness. Future challenges include continuing to improve these features and updating the content to meet the changing information needs of a rapidly evolving healthcare industry.

Acknowledgments

The authors acknowledge the significant contributions from members of the SNOMED Editorial Board, College of American Pathologists, NHS Information Authority, Kaiser Permanente CMT project, and the American Academy of Ophthalmology, including Diane Aschman, Peter Benton, Nick Booth, Phil Brown, James Campbell, Keith Campbell, Anne Casey, Colin Price, Robert Dolin, John Fedack, Paul Frosdick, Julia Kim, John Kilbourne, Debra Konicek, Karen Kudla, John Mason, Shelly Nash, Marcel Pooke, Marjorie Rallins, James Robb, Ellen Ryske, David Robinson, Joshua Schraeder, Corey Smith, and Monique van Berkum.

REFERENCES

1. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED Clinical Terms: Overview of the development process and project status. JAMIA, Symposium Supplement 2001:662-666.
2. Wang AY, Barrett JW, Bentley T, Price C, Markwell D, Spackman KA, Stearns MQ. Mapping between SNOMED RT and Clinical Terms Version 3: A key component of the SNOMED CT development process. JAMIA, Symposium Supplement 2001:741-745.
3. Kilbourne JT, Schraeder J, Smith C. Constructing a non-human subset in SNOMED CT. Proc AMIA Annual Fall Symp. In press 2002.
4. Spackman KA et al, eds. SNOMED Clinical Terms First Release. College of American Pathologists, Northfield, IL, 2002.
5. Spackman KA et al, eds. SNOMED RT. College of American Pathologists, Northfield, IL, 2000.
6. SNOMED Clinical Terms Technical Specification. College of American Pathologists, Northfield, IL, 2001. Available from: http://www.snomed.org. Accessed July 2002.
7. SNOMED Clinical Terms Technical Implementation Guide. College of American Pathologists, Northfield, IL, 2002.